Domain-Agnostic Benchmarking of Multi-Source Uncertainty for Embodied AI

1st Rebecca Keilhauer
Adaptive Autonomy & Off-road Robotics (AOR) Lab
RPTU University Kaiserslautern-Landau
Kaiserslautern, Germany
rebecca.keilhauer@cs.rptu.de

2nd Catharina Helten
Adaptive Autonomy & Off-road Robotics (AOR) Lab
RPTU University Kaiserslautern-Landau
Kaiserslautern, Germany
catharina.helten@cs.rptu.de

3rd Patrick Wolf
Adaptive Autonomy & Off-road Robotics (AOR) Lab
RPTU University Kaiserslautern-Landau
Kaiserslautern, Germany
patrick.wolf@cs.rptu.de

***Abstract*—Uncertainty is a fundamental challenge in autonomous robotics, arising from diverse sources including environmental disturbances acting on sensors, varying contextual situations, model limitations, and interaction dynamics. This paper presents a domain-agnostic benchmarking framework for multi-source uncertainty in embodied AI systems. The proposed approach captures and evaluates uncertainty at multiple levels, spanning perception, model inference, and decision-making. A set of metrics is introduced to quantify uncertainty quality, its propagation, and its impact on downstream decisions, enabling consistent comparison across systems and domains. The framework is designed to be modular and extensible, supporting heterogeneous robotic platforms and task settings. Its applicability is demonstrated in a challenging off-road autonomy scenario, in which a robotic off-road truck is exposed to multiple sources of uncertainty.**

***Index Terms*—Uncertainty Quantification, Autonomous Systems, Embodied AI, Robotics Benchmarking**

I. Introduction and Motivation

Autonomous systems often rely on multiple soft- and hardware components for environment perception, scene understanding, planning, and control execution. During operation, uncertainties can occur at every stage of this pipeline and influence the overall system behavior [1], [2]. These effects are especially relevant for safety-critical applications, since incorrect interpretations or unstable decisions can propagate through the autonomy stack and affect subsequent processing stages.

One important source of uncertainty originates from the sensor systems. Sensors, such as cameras, LiDAR, radar, GNSS, or inertial sensors, are affected by both intrinsic and extrinsic disturbances during the runtime [3]. Intrinsic sensor uncertainty originates from the sensor system itself. Measurement errors may result from sensing characteristics, limited sensor resolution, calibration inaccuracies, or measurement noise. In contrast, extrinsic sensor uncertainty arises from external environmental influences that affect sensor observations. These disturbances include illumination changes, rain, dust, reflections, partial occlusions, terrain properties, or environmental dynamics, [4]. Such effects can reduce data quality and lead to unstable observations.

Uncertainty is also introduced through contextual information. Autonomous systems frequently use additional knowledge sources, such as semantic maps, localization data, and expert knowledge. However, this information may be incomplete, outdated, or inconsistent with the current scene observations [5].

Another relevant aspect is model uncertainty within perception and decision systems. Machine learning models for object detection, classification, or semantic segmentation depend strongly on training data quality and environmental conditions [6]. Domain shifts, unseen object appearances, occlusions, or sparse sensor information can reduce prediction reliability and lead to uncertain outputs [1].

Yet, a challenge is to quantify such uncertainties during the runtime to evaluate the autonomy of a robot. Accordingly, this work presents a systematic benchmark concept for evaluating multi-source uncertainty in autonomous systems. The concept integrates uncertainty estimation methods across sensing, contextual reasoning, perception, decision-making, and control execution within a unified evaluation architecture, while standardized interfaces support interoperability between heterogeneous autonomy frameworks. The goal is to provide a benchmark applicable to arbitrary types of robots, quantifying existing uncertainties in their autonomy stack. Consequently, such uncertainties can be the triggering condition for self-adaptation of autonomy, safety evidence of engineering artifacts, or used to select relevant test criteria in scenario-based validation.

The paper is structured as follows. Section II discusses related work on uncertainty estimation in robotics. Section III introduces the uncertainty quantification methods used throughout the proposed framework. Section IV presents the benchmark conception, including the uncertainty-aware pipeline architecture and framework interfaces. Section V presents the evaluation setup, and finally Section VI concludes the paper and summarizes the main contributions.

II. Related Work

Existing approaches for uncertainty quantification in robotics mainly focus on individual layers of the autonomy

pipeline, such as perception, machine learning models, environmental context, or behavioral decision-making.

Sensor-level uncertainty is typically associated with measurement noise, sensor degradation, calibration errors, and environmental disturbances. For sensor uncertainty, the standard deviation as a quality measure can be used [2], [7]. Uncertainty estimation is important because quality-related effects originating in early perception stages can propagate through subsequent modules of the behavior network [1]. Another commonly used measure is the Signal-to-Noise Ratio (SNR) [8], which describes the ratio between signal power and noise power and can also be used as behavior control [9]. In addition, [10] estimates sensor uncertainty using Extended Kalman Filter (EKF) residuals and an adapted Mahalanobis distance to measure the statistical deviation of current sensor observations from the estimated optimal system state.

For model-based uncertainty, object detection uncertainty can be measured using the Probability-based Detection Quality (PDQ) metric, which quantifies both spatial and semantic uncertainties of detections [11]. Furthermore, [12] estimated the uncertainty region for object detection using Monte Carlo dropout.

In [13], the context uncertainty is modeled as incomplete knowledge about unobservable areas of the environment. Epistemic uncertainty is represented using spatial uncertainty regions that describe possible future positions of hypothetical objects under worst-case assumptions, such as maximum velocities and accelerations of pedestrians, bicycles, or vehicles. Furthermore, [14] modeled uncertainty in online maps for autonomous driving applications, addressing the influence of uncertainty on decision-making processes.

For behavioral uncertainty, [15] uses weighted entropy measures to influence robot behavior adaptation. In addition, steering consistency has been investigated as an indicator for behavioral uncertainty and behavioral stability [2].

[16] provides a dataset benchmark for uncertainty estimation, including sensor disturbances and object detection uncertainty. However, contextual uncertainty and uncertainty propagation across the complete autonomy pipeline are not directly considered.

Overall, most existing approaches focus on uncertainty estimation within individual layers of the robotic autonomy stack. Sensor uncertainty, model uncertainty, contextual uncertainty, and behavioral uncertainty are commonly investigated in isolation. To the best of our knowledge, no benchmark framework currently exists that systematically addresses multi-source uncertainty evaluation, including context knowledge, across the complete autonomy pipeline. Therefore, this work presents a benchmark framework concept for evaluating uncertainty across the perception-to-behavior pipeline. The proposed benchmark is intended to cover uncertainty measures at multiple abstraction levels and to enable the assessment of uncertainty propagation through perception, decision-making, and behavior adaptation processes.

III. Uncertainty Quantification

This section describes the uncertainty quantification methods used at different levels of the autonomy pipeline. Uncertainty is modeled across sensing, contextual reasoning, perception, classification, decision-making, and action execution (see Fig. 1).

Level	Function	Description
Sensor	σ_{sensor}	Sensor measurement uncertainty
Context	H_{Co}	Semantic context uncertainty
Context	σ_{context}	Geospatial/contextual uncertainty
Perception	H_{B}	Bounding box uncertainty
Perception	σ_{B}	Bounding box temporal uncertainty
Perception	H_{Cl}	Classification uncertainty
Control	H_{D}	Decision uncertainty
Control	σ_{action}	Action control uncertainty

Fig. 1. Uncertainty Quantification Methods.

A. Sensor Data Uncertainty

Sensor uncertainty describes the variability of raw sensor measurements obtained from systems such as cameras, LiDAR, radar, or inertial sensors. The uncertainty is quantified using the standard deviation σ of the observed measurements [2], [7]. A high standard deviation indicates unstable or noisy measurements, while a low standard deviation corresponds to more consistent sensor observations.

$$\sigma := \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \tag{1}$$

where μ represents the mean value of the sensor measurements with cardinality N [17]. Sensor data uncertainty originates from intrinsic uncertainties caused by the sensor model itself, as well as extrinsic uncertainties caused by environmental influences such as weather conditions, illumination changes, dust, terrain properties, or sensor occlusions.

B. Context Uncertainty

Context uncertainty represents uncertainty within semantic and geospatial contextual information. Two complementary measures are used for its quantification. Semantic context uncertainty is modeled using entropy:

$$H := -\sum_{x\in\mathcal{X}} p(x)\log p(x), \tag{2}$$

where $\mathcal{X}$ represents the set of possible states and $p(x)$ denotes the probability of state x [17]. For the semantic context, we denote the entropy by H_{Co}. Higher entropy values indicate increased ambiguity within the contextual interpretation, whereas lower entropy corresponds to more certain contextual knowledge. For geospatial or numerical contextual information, uncertainty is quantified using the standard deviation of Eq. 1, denoted here as σ_{context}

C. Bounding Box Uncertainty

Bounding box uncertainty describes the uncertainty associated with object localization and spatial representation within the perception system. The ambiguity of bounding box predictions is quantified using entropy (see Eq. 2) denoting as H_{B}, , whereas temporal or spatial fluctuations of bounding box estimates are quantified using the standard deviation σ_{B}.

D. Classification Uncertainty

Classification uncertainty quantifies ambiguity in semantic class predictions. Entropy is used to measure the uncertainty of the predicted class probability distribution, denoted by H_{Cl} (see Eq. 2).

E. Decision Uncertainty

Decision uncertainty describes uncertainty in behavior selection and planning decisions. The uncertainty is quantified using entropy over the probability distribution of possible decisions, denoting here as H_{D}. A high entropy value indicates uncertain decision-making with multiple competing alternatives, whereas low entropy corresponds to confident decision selection.

F. Action Control Uncertainty

Action control uncertainty quantifies the stability and consistency of executed control commands such as steering angles, acceleration, or velocity outputs. The uncertainty is estimated using the standard deviation (σ_{action}) of the control signals (see Eq. 1). Large standard deviation values indicate unstable or inconsistent control behavior, while smaller values correspond to stable and smooth action execution.

IV. Benchmark Conception

Fig. 2 illustrates the proposed benchmark architecture for autonomous systems with integrated uncertainty measurement. The concept considers uncertainty propagation across the autonomy pipeline, ranging from sensor measurement to action execution. The architecture combines sensing, contextual reasoning, perception, decision-making, and control within a modular evaluation structure. The pipeline consists of two primary information streams: sensor data and context data. Sensor data includes observations from systems such as cameras, LiDAR, radar, GNSS, and inertial sensors. Since these observations are affected by measurement noise and environmental influences, uncertainty estimation is performed directly after data acquisition. Sensor uncertainty is quantified using the standard deviation σ_{sensor}.

Context data includes environmental knowledge such as map information, localization data, terrain descriptions, or expert knowledge. The benchmark distinguishes between numerical and symbolic contextual uncertainty. Numerical context uncertainty is represented using the standard deviation $\sigma_{\mathrm{context}}$, while semantic ambiguity is quantified using entropy H_{Co}.

After uncertainty estimation, the sensor data are forwarded to the perception system. The perception layer may contain machine learning models, classical computer vision methods, or sensor fusion approaches generating scene representations such as 2D and 3D bounding boxes and semantic classifications.

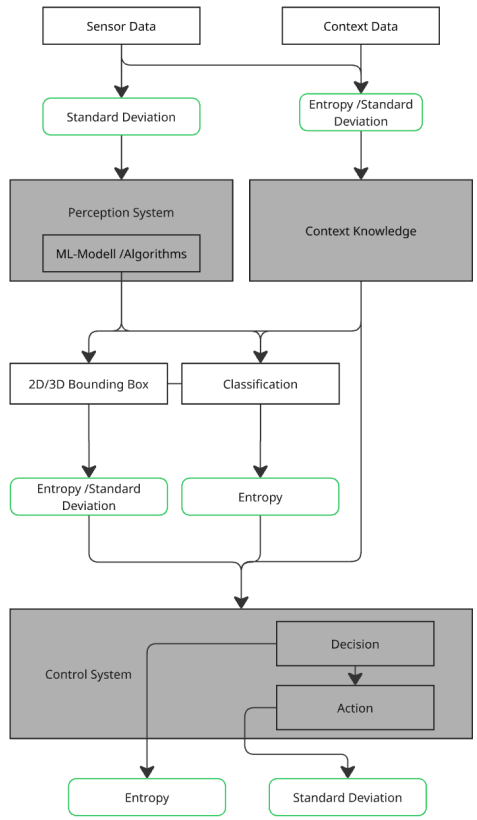

Fig. 2. Autonomous robotic perception and control architecture with integrated uncertainty estimation.

Bounding-box uncertainty is quantified using entropy-based uncertainty estimation H_B. In addition, temporal and spatial fluctuations across consecutive observations are quantified using the standard deviation σ_B. Classification uncertainty is quantified using entropy H_{Cl} over the predicted class probabilities.

The benchmark additionally incorporates contextual knowledge into the perception process. Context information can reinforce plausible detections and suppress inconsistent interpretations. For example, classifications that contradict the environmental context may produce increased uncertainty estimates.

The resulting perception outputs are forwarded to the control system, consisting of a decision module and an action module. Decision uncertainty is represented using entropy H_{D} over the decision probability distribution. Action uncertainty is quantified using the standard deviation σ_{action} of executed control commands such as steering, acceleration, or velocity outputs.

The benchmark, therefore, enables the systematic analysis of uncertainty propagation across all autonomy layers, including sensing, contextual reasoning, perception, semantic interpretation, decision-making, and control execution.

A. Framework Interfaces

To ensure interoperability across heterogeneous robotic software stacks, the benchmark defines abstract interfaces between all pipeline components independently of specific middleware, autonomy frameworks, or hardware platforms. Perception outputs are represented using standardized 2D and 3D bounding-box interfaces containing class labels, spatial positions, dimensions, timestamps, confidence scores, and uncertainty values. Context interfaces provide environmental and semantic information independently from the underlying map representation or reasoning framework Decision interfaces represent probabilistic behavior outputs and planning hypotheses including confidence values for candidate actions. Action interfaces encode low-level control commands such as steering, velocity, acceleration, and actuator targets together with associated uncertainty measures. Overall, the modular interface structure enables the integration of different robotic software architectures and autonomy systems within a unified uncertainty evaluation framework.

V. Evaluation

The following provides an excerpt of benchmark results. Therefore, the test environment, robot, and autonomy stack are introduced, with exemplary uncertainty results.

A. Autonomy Framework and Test Environment

The autonomy approach that serves as a demonstrator for the uncertainty evaluation uses the REACTiON framework, as proposed by [2], and was analyzed by the benchmark. Therefore, the corresponding uncertainty quantification mechanisms have been implemented and integrated into the autonomy software. The robot's autonomy stack itself was implemented using the C++-based middleware FINROC [18].

The evaluation and benchmark use data recording from a landfill [19]. The landfill consists of a former landfill service track used as an off-road navigation scenario. The route contains degraded terrain structures caused by erosion and vegetation growth. The environment includes uneven off-road surfaces, vegetation-covered paths, trenches, mud holes, deep furrows, and steep terrain inclinations. Additionally, the environment includes nearby asphalt roads, parking areas, bushes, and surrounding vegetation.

Within the landfill, an autonomous Unimog truck followed a rugged pathway. The vehicle is a Unimog 5023 equipped with a front-facing stereo camera (ZED), SICK LMS 511 laser, a Robosense RS-LiDAR-16 laser scanner, a Microstrain 3DM-GX5-25 IMU, and a u-blox NEO7P GNSS system. The vehicle itself is highly off-road capable, with controllable differentials and a gearbox [2].

B. Excerpt of the Uncertainty Benchmark Results

Exemplary benchmark results are depicted in Fig. 3. Here, the sensory uncertainty directly corresponded to the traveled terrain. If the robot tilted during off-road travel (e.g., at time points 25 and 325), the uncertainty decreased because the sensors covered the ground in front of the robot. Otherwise, if only faraway regions are sensed, the uncertainty about the precision of the sensory measurements increases. Similarly, the other uncertainty factors (cf. Section 3) can be tested, such as context and bounding boxes (see Fig. 4)

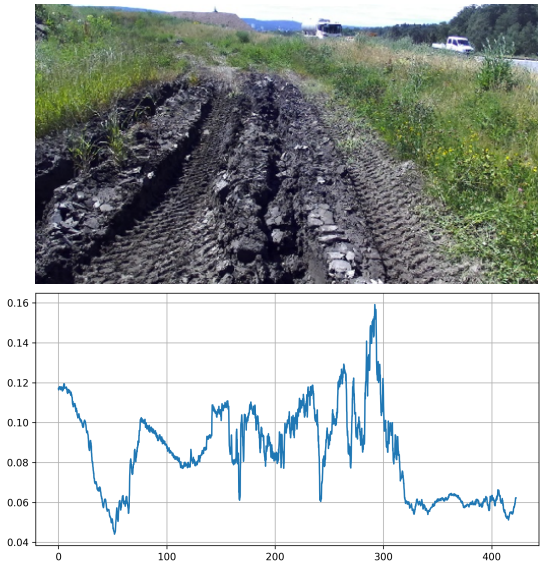

Fig. 3. Laser point cloud average sensor data uncertainty while traveling the landfill.

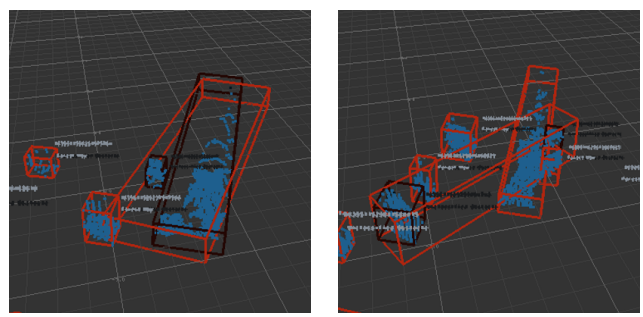

Fig. 4. Fluctuations in object classification and clustering between two consecutive time frames lead to an increase in contextual uncertainty represented by entropy.

VI. Conclusion

This paper presented a domain-agnostic benchmark framework for uncertainty evaluation in autonomous systems. The proposed concept integrates uncertainty estimation across sensing, contextual reasoning, perception, decision-making, and control execution within a modular architecture with standardized interfaces.

Currently, the benchmark mainly represents a conceptual framework. The uncertainty measures are considered independently, without directly modeled mathematical relations or a global uncertainty metric. Furthermore, the evaluation was limited to a set of data recordings and a single robot navigating in off-road areas.

Future work will focus on mathematically grounded uncertainty propagation models, real-time evaluations, and validation on different robotic platforms and autonomous systems.

References

[1] P. F. Pelz, M. E. Pfetsch, S. Kersting, M. Kohler, A. Matei, T. Melz, R. Platz, M. Schaeffner, and S. Ulbrich, *Types of Uncertainty*. Cham: Springer International Publishing, 2021, pp. 25–42. [Online]. Available: https://doi.org/10.1007/978-3-030-78354-9_2

[2] P. Wolf, *Cognitive Processing in Behavior-Based Perception of Autonomous Off-Road Vehicles*. Munich, Germany: Verlag Dr. Hut GmbH, October 10 2022.

[3] P. Wolf and K. Berns, "Data-fusion for robust off-road perception considering data quality of uncertain sensors," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Prague, Czech Republic, September 27–October 1 2021, pp. 6853–6860.

[4] R. Heinzler, P. Schindler, J. Seekircher, W. Ritter, and W. Stork, "Weather influence and classification with automotive lidar sensors," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE Press, 2019, p. 1527–1534.

[5] P. Wolf, A. Vierling, T. Ropertz, S. Velden, C. Guzman, and K. Berns, "Autonomous off-road navigation using near-feature-based world knowledge incorporation on the example of forest path detection," in *16th International Conference On Intelligent Autonomous Systems (IAS)*, Singapore, June 22–25 2021.

[6] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521001081

[7] T. Ropertz, P. Wolf, and K. Berns, "Quality-based behavior-based control for autonomous robots in rough environments," in *Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2017)*, O. Gusikhin and K. Madani, Eds., vol. 1. Madrid, Spain: SCITEPRESS – Science and Technology Publications, Lda, July 26–28 2017, pp. 513–524.

[8] E. Schrüfer, L. Reindl, A. König, and B. Zagar, *Elektrische Messtechnik: Messung elektrischer und nichtelektrischer Größen*. Carl Hanser Verlag GmbH & Company KG, 2012. [Online]. Available: https://books.google.de/books?id=3alPAgAAQBAJ

[9] W. Wu and F. Zhang, "Experimental validation of source seeking with a switching strategy," 05 2011, pp. 3835–3840.

[10] M. Alharbi and H. Karimi, "Context-aware sensor uncertainty estimation for autonomous vehicles," *Vehicles*, vol. 3, pp. 721–735, 10 2021.

[11] D. Hall, F. Dayoub, J. Skinner, H. Zhang, D. Miller, P. Corke, G. Carneiro, A. Angelova, and N. Sünderhauf, "Probabilistic object detection: Definition and evaluation," 2020. [Online]. Available: https://arxiv.org/abs/1811.10800

[12] L. Peng, H. Wang, and J. Li, "Uncertainty evaluation of object detection algorithms for autonomous vehicles," *Automotive Innovation*, vol. 4, 07 2021.

[13] M. Nolte, S. Ernst, J. Richelmann, and M. Maurer, "Representing the unknown – impact of uncertainty on the interaction between decision making and trajectory generation," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2412–2418.

[14] P. Yang, Y. Zheng, Q. Zhang, K. Zhu, Z. Xing, Q. Lin, Y.-F. Liu, Z. Su, and D. Zhao, "Uncad: Towards safe end-to-end autonomous driving via online map uncertainty," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2025, p. 6408–6415. [Online]. Available: http://dx.doi.org/10.1109/ICRA55743.2025.11128148

[15] A. Suresh, C. Nieto-Granda, and S. Martínez, "Robotic exploration using generalized behavioral entropy," *IEEE Robotics and Automation Letters*, vol. 9, no. 9, p. 8011–8018, 2024. [Online]. Available: http://dx.doi.org/10.1109/LRA.2024.3433207

[16] G. Franchi, X. Yu, A. Bursuc, A. Tena, R. Kazmierczak, S. Dubuisson, E. Aldea, and D. Filliat, "Muad: Multiple uncertainties for autonomous driving, a benchmark for multiple uncertainty types and tasks," 2022. [Online]. Available: https://arxiv.org/abs/2203.01437

[17] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambrigde, Masschussetts, London, England: MIT Press, 2006.

[18] M. Reichardt, S. Schütz, and K. Berns, "One fits more - on highly modular quality-driven design of robotic frameworks and middleware," *Journal of Software Engineering for Robotics (JOSER): Special Issue on Robotic Computing*, vol. 8, no. 1, pp. 141–153, December 2017, iSSN: 2035-3928; URL: http://joser.unibg.it/index.php?journal=joser&page=issue&op=view&path[]=9.

[19] P. Wolf and K. Berns, "Dataset: Extreme uphill to tent: Narrow path navigation," 2020. [Online]. Available: https://doi.org/10.26204/data/18